Effectiveness of group interaction on conceptual standardized test performance

Chandralekha Singh, University of Pittsburgh

We analyse the effectiveness of working in pair on the Conceptual Survey of Electricity and Magnetism test in a calculus-based introductory physics course. We discuss the implications of pairing students with different individual achievments.

1 Introduction

Peer collaboration has been exploited as a learning tool in many diverse instructional settings and with different types and levels of student populations. Here, we analyse the effectiveness of working in pairs on the performance on a standardized conceptual multiple-choice test, Conceptual Survey of Electricity and Magnetism (CSEM), in a calculus-based introductory physics course. In two consecutive semesters, in double class periods (1 hour 50 minute stretch), students were administered the test individually and in groups of two after instruction of the relevant concepts. Each time (whether group or individual) students were allowed 50 minutes to take the CSEM test. Between the individual and group administrations of the test there was a short 5-10 minute break and students were required to turn in their first response so that they could not refer to it when working in group. The test answers were never discussed with students so when they switched from individual to group (or vice versa), they did not know if their initial responses were correct or not.

Although some studies show that heterogenous groups are more effective for group learning, others show that working with friends has special advantages. In this study, students were allowed to choose their own partners. They were encouraged to discuss the response with each other, and each test counted for one quiz grade. Students had an additional incentive to discuss the concepts because an examination in two weeks covered the same material. Moreover, both semesters, students had extensive experience working with two peers in the recitation on context-rich problems and with one peer during the lecture on Mazur-style concept tests.

We note that the peer collaboration was unguided in that there was no help or facilitation from the instructor. Therefore, the trends that emerge may be applicable to students working together outside of the class but is likely to be different from those in collaborations between a tutor and tutee where one person's knowledge is on significantly firmer ground. One attractive feature of peer collaboration is that since both peers have recently gone through similar difficulties in assimilating and accomodating the new material, they can often relate to each other's difficulties more easily than the instructor. The instructors' extensive experience can often make a concept so obvious and automatic that they may not comprehend why students are misinterpreting various aspect of a concept or finding them problematic and confusing.

2 Discussion

The test was only administered after the instruction of the relevent concepts because our goal is to assess the effectiveness of group dynamics (not overall instruction). To obtain two random equivalent samples, all students in the class sitting on one side of the isle took the individual test first followed by the group test (IG treatment: 74 students or 37 pairs) while those on the other side of the isle took the group test first before taking it individually (GI treatment: 54 students or 27 pairs). One reason for giving the test in both orders was to assess the effect of thinking individually before the peer discussion. In the Mazur-style peer instruction, students are first asked to think about the concepts based on the as-

sumption that not allowing an opportunity to think individually may prevent students from evaluating their own stand on a concept. Another reason for designing both the IG and GI treatments was to evaluate the "test-retest" or "practice" effect.

Although the trends on some individual test items are interesting, here we only discuss the effect of group work on the overall test scores due to the space limitations. In the GI treatment the average group score was 71.7% compared to 70.3% average on individual test that followed. Thus, on the average, students performed the same in the group and in the individual testing that followed immediately (It should be noted that students could not refer to their group work when they worked individually). In the IG treatment, the average group score was 72.5% compared to the average individual score of 55% (normalized gain 0.39). The fact that the group performance on the IG and GI treatments are virtually indistinguishable (72.5% vs. 71.7%) suggests that giving students an opportunity to think individually before the peer discussion did not improve their group performance. Also, in the IG treatment, the normalized gain of 0.39 in the group work is clearly not a "test-retest" effect because considering the treatment samples to be equivalent, we can compare the individual performance on IG treament (55%) with the group performance of GI treatment (71.7%) which shows a gain of 0.37 (indistinguishable from 0.39). Therefore, the rest of the discussion will mostly be focussed on the IG treatment.

But before moving on, we note some interesting trends in the amount of time students took in the IG and GI treatments during the two testings in immediate succession. In the GI treatment, during the group work and in the IG treatment, both during the individual and group work students used the same amount of time. One the other hand, students working individually after the group work in the GI treatment on an average took only about one third of the initial time spend on group work. It appears that in the IG treatment, despite having worked on the problems individually, students were willing to spend the time discussing the same test because they found the peer discussion useful. On the other hand, in the GI treatment, after having discussed the test with peers, students were reasonably sure about their thoughts and did not consider it necessary to brood over the problems again.

2.1 Evidence for co-construction

Although there is no concensus in the research literature on the definition of "co-construction", we use this term to denote cases where neither student alone chose the correct response but the group discussion helped converge on the correct response. Co-construction can occur for several reasons. For example, if the group memebers chose *different* incorrect responses, they will have to explain their reasoning to each other. This can unravel problems in their initial logic and complementary information provided by peers can help them converge on the correct solution. Even in cases where both students have the *same* incorrect response, co-construction can occur if students are unsure about their initial response and are willing to discuss their apprehensions with peers. Important clues provided by peers during the discussion can trigger recall of revelant concepts and can help the group co-construct.

Table 1 displays the fraction of pair of correct-incorrect response patterns based upon students' individual response and how they changed in the group work (for IG treatment). It shows evidence for co-construction in 7.5% of overall cases.

Ind. Resp.	Group Resp.	
00 (0.26)	000 (18%)	001 (7.5%)
01 (0.38)	010 (8.5%)	011 (30%)
11 (0.36)	110 (0%)	111 (36%)

Table 1: 0 *refers to an incorrect response and* 1 *refers to a correct response. The fractions associated with* 00 *(both incorrect),* 01 *(one incorrect), and* 11 *(both correct) is based upon the individual response across all items while the percentage for* 000*, for example, refers to the group response in which both individuals and the group had incorrect response.*

To verify that the 001 cases in which both students individually chose incorrect response but the group chose the correct response is not due to "just guessing", we

analyse the first row of table 1 in detail. Table 2 subdivides this row based upon whether both partners had the same or different incorrect responses and if the group response was one of the original incorrect response or a third incorrect response.

Ind. Resp.	Group Resp.		
	1	0 or $0''$	$0'$
00 (same incorr.)	22%	70%	8%
$00''$(diff. incorr.)	34%	52%	14%

Table 2: *Distribution of group response for the cases where both peers had the same or different incorrect responses.* 0, $0'$ *and* $0''$ *refer to different incorrect responses.*

Table 2 shows that for 00 (same incorrect), in 22%, and for $00''$ (different incorrect), in 34% of the cases, group response was correct. In comparsion, the relatively small frequency of the incorrect group responses (see table 2) that were not originally selected by either individuals suggests that students were not "just guessing". Although we did not conduct formal interviews with students after they worked in groups, we briefly discussed the aspects of the group work that were helpful with several students. Most students admitted that they got useful insights about various electricity and magnetism concepts by discussing them with peers. Students frequently noted that they sometimes had difficulty interpreting the problems alone but it became easier with a friend. They also said that talking to peers forced them to organize their thoughts, find fault with their initial reasoning, and reminded them of concepts they had difficulty recalling on their own. Also, qualitative observation shows that students were more likely to draw field lines, write equations or scribble on their exams in the group work than in the individual work.

2.2 Negative impact of grouping?

Table 1 shows that out of all the 01 cases in which one student individually had the correct response and the other had an incorrect response, 78% of the group responses were correct. The fact that 22% of 01 resulted in incorrect group response (010) is only slightly alarming because it was an unguided peer discussion. It can happen if students who individually chose the correct response are not very confident and cannot defend or justify their response. However, the fact that a majority of 01 resulted in 011 suggests that students who individually chose the correct response were generally more confident and were able to justify their choice. It may also suggest that students who got the item wrong individually were unsure of their choice and were willing to listen to their peers. As we will see in the next section, group work never resulted in a negative individual gain.

2.3 Individual gain and retention

One can hypothesize that the *individual* performance in the GI treatment was much superior (70.3%) compared to the IG treatment (55%) because students could immediately recall the group responses for all the 32 test items in the GI treatment and their superior performance does not reflect the effectiveness of group work with regard to the retention of the concepts discussed. Similarly, in the IG treatment, the superior group performance compared to the individual performance is due to a large number of 011 but it does not mean that the student who got 0 individually will retain what they learned in the group work. To test this hypothesis, two weeks after the IG and GI treatments, all students individually took the CSEM test again. In view of the fact that students earlier went through either the IG or GI treatments, we now relabel the treatments IGI and GII. We note that students did not know that they will be taking the same test again. Although it would have been significantly better if another equally reliable test for assessing similar concepts could be used; its unavailability lead us to use the same test. However, the test was never discussed with students at any time. We are currently conducting a study in which students individually took the test once and then again two weeks later *without* any group intervention to tease apart the effectiveness of group work vs. the effect of having seen the test and having the opportunity to study before the second individual testing.

In the IGI treatment, the average for the second individual testing was 74% which is a gain of 0.42% compared to the initial individual score of 55%. A detailed comparison of the scores for the group vs. the sec-

ond individual adminstration shows that 81% of the overall individual responses chosen by the members of a particular group were the same as the group responses. Out of the 19% (second) individual responses that were different from the group, roughly 11% went from incorrect to correct while 8% went from correct to incorrect. The corresponding numbers in the GII treatment are virtually the same. This suggests that the gist of group interaction was mostly retained even after two weeks.

2.4 Effective pairing

To learn about the type of pairings that will optimize the overall gain for this conceptual test, we divided the 75 students in the IGI treatment into three groups: I (high), II (middle), and III (low), based upon their initial individual score on CSEM. In tables 3a and 3b below, we show the average initial individual score (left) and the normalized gain in the second individual testing (right) for the 9 types of pairs: (I I), (I II), (I III), (II I), (II II), (II III), (III I), (III II), and (III III). The top row of the table refers to the performance of type I students for different kinds of pairings: (I I), (I II), and (I III). Similarly, the middle and bottom row refer to the performance of type II and type III students respectively. For comparison, for all 75 students together, the average initial individual score was 55% and the average gain was 0.42.

	I	II	III
I	73	74	71
II	56	54	56
III	34	36	36

I	II	III
0.54	0.55	0.29
0.65	0.51	0.26
0.41	0.40	0.37

The above table shows that although all students in the IGI treatment gained in the second individual testing two weeks later compared to the first, the gain matrix is not symmetric. For example, the gain of I due to the interaction with III (0.29) is not the same as that of III due to I (0.41) in the (I III) pairing. It is striking that the gain of type I and type II students is significantly lower when they paired with type III students. Type I students have similar gains whether they paired with type I or type II students while Type II students benefit more from pairing with type I than with another type II student. Interestingly, the gain of type III students (lowest third) is virtually the same regardless of who they paired with (bottom row). One hypothesis is that while pairing with a higher achievement student provided type III with an opportunity to do well in the group test, they did not retain all of the concepts because they were subdued by the higher achieving student and did not participate very actively in the discussions. Thus, the opportunity to learn from the higher achieving student may have been outweighed by the inability of type III to process the information at the rate discussed by the other student and their inability to participate fully in the discussions which is crucial for retention. On the other hand, in the (III III) pairing, both students had comparable but some complementary knowledge and both actively participated in the discussions. The evidence for this hypothesis comes from the comparison of the average group (left) and second individual test score (right) for each of the pairings as shown in tables 4a and 4b below:

	I	II	III
I	88	85	72
II	85	73	62
III	72	62	54

I	II	III
88	88	79
85	78	67
61	62	60

A comparison of tables 4a and 4b shows that the average individual score of type III students in the (I III) pairing after two weeks deteriorated (61%) compared to their group score (72%). Also, a comparison of tables 3a and 4a shows that type I students did not benefit from interactions with type III students and the average initial score for type I students and the group score for the (I III) pairing are the same. Similar comparisons for type II students shows that although they benefitted from all types of interactions their gain improved as they interacted with higher achieving students. It appears that at least for this conceptual test, the pairing that helps maximize the overall gain is one that only has (I II) and (III III) types of pairs. It will be useful to investigate the extent to which this result is universal, i.e., two peers collaborating on conceptual tests show highest overall gains when the high and middle achievement students are paired and the low achievement students are paired with each other.